Pedselect

Large multiplex pedigrees can be informative for helping identify disease associated risk loci. As many pedigrees are collected over long periods of time, pedigree members may vary in terms of whether their phenotype has been observed, whether they have already been sequenced/genotyped, and whether they have DNA collected that can readily be sequenced/genotyped. Given budget constraints, only sequencing some members of the pedigree may be feasible. Pedigree-based genotype imputation methods (e.g. Cheung et al., 2013) can then be used to 'fill-in' the genotypes of members who have not (or cannot) been sequenced.

Pedselect is a tool for selecting the most informative individuals to perform whole-genome-sequencing/genotyping in a pedigree. In order to maximize power to identify disease associated loci, Pedselect first prioritizes members whose phenotypes have been collected and have DNA available. It then sequentially selects members to sequence based on their ability to inform the genotypes of nearby relatives. Once all individuals with phenotypes and available DNA have been selected, Pedselect then sequentially selects among the remaining individuals who have DNA available but have not been phenotyped based on their ability to inform the genotypes of remaining phenotyped individuals (who do not have DNA available).

Method

In order to quantify "imputability", we derived a score for each individual *i* for whose genotypes we wish to impute: $S_i = \sum_{j=1}^{n} \frac{1}{m_j}$, where $j = 1,2,\ldots n$ are the genotyped relatives of individual *i* and *m* is the number of meiosis events, or nodes, between individual *i* and *j* (see Figure 1). When calculating S_i, we do not count a genotyped distant relative if a closer genotyped relative along the same relatedness path is observed. For example, if the genotypes of both the mother and maternal grandmother are available, we do not consider information from the grandmother since all shared identical-by-descent segments with the grandmother are also observed in the mother (Rafnar et al., 2011).

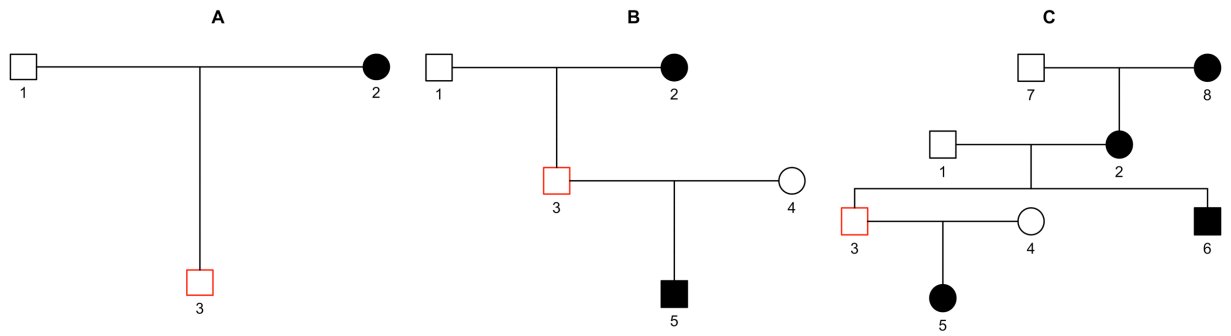

Figure 1. Three examples for calculating the "imputability" score from pedigrees. Filled-in individuals are those with existing genotype information. The individual in red is who we wish to impute. A) Here, the imputability score for the red individual = 1, since only the genotypes of his mother (1 node) can be used for imputation. B) Here, the imputability score for the red individual = 2, since information from his mother (1 node) and child (1 node) can be used for imputation. 3) Here, the imputability score for the red individual = 2.5, since information from 3's mother (1 node), child (1 node) and sibling (2 nodes) can be used for imputation (score = 1/1 + 1/1 + 1/2). We do not consider the his grandmother since all shared IBD segments can be observed in the mother.

We use this imputability score to prioritize which individuals with available DNA and a known phenotype to sequence. Our algorithm goes through each of these individuals in turn and, assuming that the individual is sequenced, calculates the imputability score of the remaining individuals who have phenotypes but no genotypes. We then select for sequencing the individual who maximizes the sum of the imputability scores of the remaining individuals. This procedure repeats until no phenotyped individuals who can be sequenced remain. The procedure then starts again, this time selecting among unphenotyped individuals who have available DNA, and prioritizing them based on whether have their genotypes can inform the genotypes of additional phenotyped (but with no available DNA) individuals.

Intuitively, the final list of individuals to sequence is based on whether they have a known phenotype themselves and on their ability to inform the genotypes of additional phenotyped individuals in the pedigree.

Pedigree simulations

To illustrate the relationship between the imputability score and pedigree imputation performance, we simulated 1000 outbred pedigrees of random size (range 11 to 177). Genotypes for a single chromosome with 10,000 bi-allelic SNPs were generated for all individuals in each pedigree using

SimPed (Leal et al., 2005). We set the genetic distance between each SNP to be 0.01cM. The allele frequency of each SNP was randomly assigned so that the distribution of allele frequencies closely resembles those of the 1000 Genomes Phase 3 EUR samples. For each pedigree, we randomly selected one individual to be the target individual to be imputed and masked their genotypes. We then set to missing the genotypes of a random number of other individuals within each pedigree (range 2 to 168). Hence, for each randomly generated pedigree, we use only a subset of individuals who have a known genotype to calculate the imputability score and impute the genotypes of the target individual.

Pedigree imputation was performed using GIGI (Cheung CYK et al., 2013). For each pedigree, we selected a set of common markers (minor allele frequency > 0.05) separated by a distance of at least 0.5cM to be the framework panel used by GIGI. The number of framework markers per pedigree ranged from 463 to 473. After imputation, genotypes were hard-called if the imputed genotype probability > 0.95 or missing otherwise. Overall, we see a strong positive relationship between the proportion of SNPs successfully imputed and the imputability score (Figure 2).

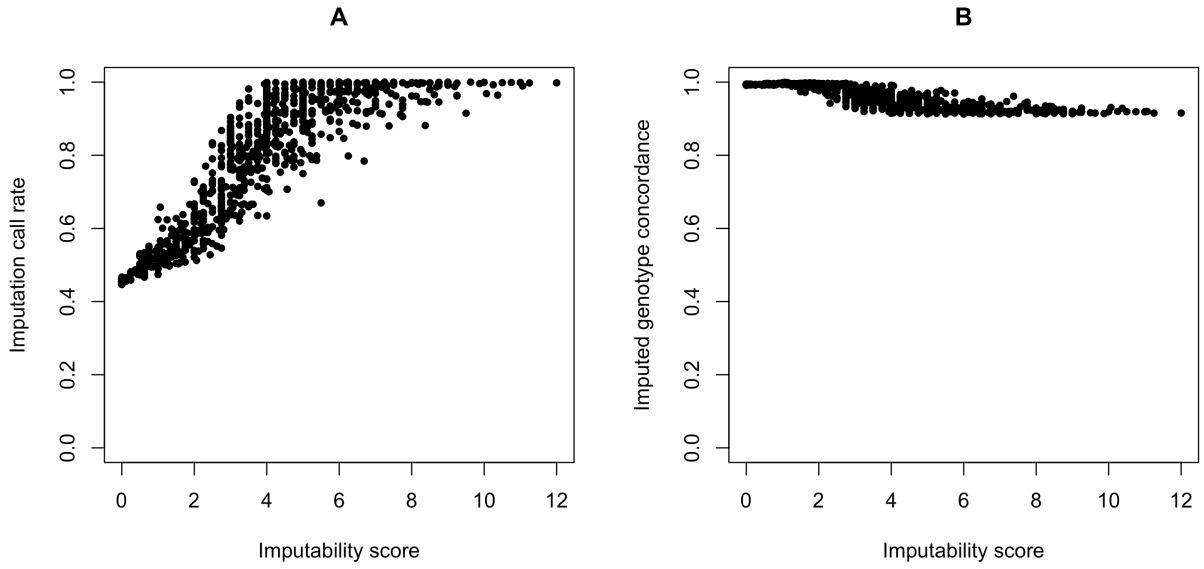

Figure 2. Relationship between imputability score and imputation performance from 1000 simulated pedigrees as measured by (A) call rate and (B) genotype concordance. Note that even at a score with 0 (i.e. no close relatives of the target individual has genotypes available), GIGI was able to impute ~45% of SNPs. These almost all reflect rare variants that GIGI called on the basis of allele frequencies and passed the genotype probability threshold of 0.95.

References

Cheung CYK, Thompson EA and Wijsman EM, GIGI: An approach to effective imputation of dense genotypes on large pedigrees. Am J Hum Genet. 2013,91:504-16

Leal SM, Yan K and Müller-Myhsok B, SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. Hum Hered. 2005,60:119-22

Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A et al., Mutations in BRIP1 confer high risk of ovarian cancer. Nat Gen. 2011, 43:1104-7

Contact

Jimmy Liu, New York Genome Center (first name dot z dot last name at gmail dot com)

March 30 2017